Werner Güth and Hartmut Kliemt

From full to bounded rationality

The limits of unlimited rationality

Abstract: Deriving advice that can in fact be utilized by boundedly rational decision makers is a central function of modeling choice making. We illustrate why this role is not being fulfilled well by standard models of full rationality and that theories of bounded rationality are needed not only for better predictions, but also for developing better advice. Our main point is that one cannot succeed here without studying how theories of bounded rationality causally influence the behavior of boundedly rational individuals. In view of such a causal role of theories we discuss how advice of a theory of boundedly rational behavior can become known, be followed among boundedly rational individuals and still be good advice.

1. Explications and explanations

Ideally the theoretical concepts of social science should be precise, fruitful, simple and similar to the pre-theoretical ones which they substitute. The process in which a theoretical substitute for a pre-theoretical concept is worked out – as well as the result of this process – is called an „explication“ (see on the concept of an explication originally Carnap 1956). An explication has descriptive as well as normative aspects. On the one hand, the general use of the pre-theoretical concept must be described, on the other hand, a more precise, theoretically fruitful, yet simple and similar concept must be formed and established as a „standard“.

In the so-called „moral sciences“, explicating the concept of „rationality“ is one of the most important challenges. First, „rational” and „irrational” are prominent evaluative classifications in daily as well as scientific life. This renders the task

of explicating them an important one. Second, the actual use of the term „rational“ varies so much that even the most careful description will not reveal a common conceptual core.

Avoiding part of the challenge, traditional rational choice theorists have not paid much attention to the everyday or common use of the term „rational“ or to actual behavior. They established their standards of rationality – thus fixing the meaning of the term „rational” – independently of or even counter to the facts. For instance, even if most individuals used the term „irrational“ for characterizing polluting behavior in an n-person „public bads’” experiment, the typical rational choice theorist would tend to classify the participants’ behavior as „rational”. As a matter of fact, if many people should say that individuals who do not co-operate in a one-off classical two-person prisoners’ dilemma situation are behaving „irrationally“, this will not make the rational choice theorist think twice.

If philosophers like Edward McClennen (1990) insist that „backward induction“ in the finitely repeated prisoners’ dilemma must be given up since it just cannot be rational *not* to co-operate at least to some extent in such situations, the adherent of theories of perfect rationality will remain unmoved. Her concept of „full rationality” is formed „deductively” rather than „inductively” and justified by a priori normative rather than a posteriori descriptive reasoning. Common usage of the terms „rational“ and „irrational“ as well as philosophical analyses starting from common intuitions seem simply misguided to the traditional rational choice theorist. She would claim that neither actual behavior nor actual usage of terms should determine the „proper“ meaning of the term „rational“ in a scientific or philosophical context.

The „a priori” approach is clearly a possibility. But it must be noted, too, that by establishing standards of rationality independently of or counter to the facts, the rational choice theorist turns „rationality“ into what may be called a „counterfactual concept“. If we take such an approach to its extreme, we get a very refined rationality-concept that may be appealing to the theorist, in particular the decision and the game theorist. However, it neither relates directly to the understanding of real people, nor does it apply to their behavior, nor can it form

a general basis for formulating advice that is helpful for merely boundedly rational individuals.

It becomes even doubtful whether there is still sufficient continuity between the pre-theoretical and the theoretical concept of rationality to use the same term for both. Since the continuity issue must not be taken lightly, the question arises of whether there may be more „realistic“ explications of the rationality concept in terms of so-called „bounded rationality”. As in the somewhat parallel case of „perfect“ and „workable competition“, such explications would uphold some essential elements of the purely theoretical „normative concept“ (normative at least in the sense of an ideal type), while moving the concept closer to real behavior in other regards.

Construing a reflective equilibrium as to the proper meaning of „rationality“ within the concept of „bounded rationality“ raises fundamental questions of a very general nature. We shall address such questions subsequently in an intuitive way by discussing quite trivial examples borrowed from game theory. But we do not intend to play by the rules of the traditional game theoretic discourse in which acceptance of some extreme assumptions is required from all participants. Rather, we are willing to challenge these assumptions whenever necessary or suitable.

In the next section, the relation between theory and advice is discussed in a general manner (2.). Then the decision environment on which we shall focus for illustrative purposes is introduced (3.). After specifying more formally the concept of advice (4.), the relation between knowing and complying with advice is discussed (5.). In the main section, responses to advice are studied (6.). In the subsequent section a „causal” rather than „logical” interpretation of the role of advice as behavioral guidance is suggested and related to the concept of a semi-normative theory (7.). Some general observations on the continuity between pre-theoretical and theoretical explications of concepts of rationality bring the argument back full circle and conclude the paper (8.).

2. Theory and advice

It seems ridiculous to claim that conventional game theory is a theory of how real people interact. It is „descriptive“ merely in the sense of characterizing what *would* happen if inhabitants of a world of fully rational beings *were* to interact rationally with each other. In this hypothetical world – at least in a way – the normative perspective coincides with the descriptive one since – by assumption – rational actors do what rational actors ought to do. Advice as to what rational actors ought to do in view of the rationality ascribed to them coincides with predictions of what they will do on behalf of their rationality. What this will be, of course, can be derived only after the rules of the game in question have been specified. Once those rules are fixed, a closed theory of rational play would offer a suggestion or an advice of how the players should – and would – play. Under the assumption that it be completely followed by all players concerned, such a closed theory can eliminate the problem of strategic uncertainty by providing completely specified, unambiguous advice of how to play (see Harsanyi and Selten 1988).

The practical value of advice to a world of rational beings is quite precarious in a world of less than fully rational individuals. As in second-best theory, where removing some obstacles to efficiency can lead to yet inferior results if other obstacles remain in place, approximating ideal behavior in the real world may not be the best strategy. In fact, criteria for what is good advice in the real world are rather unclear. Although arriving at more clarity here is quite difficult, some hints can be given at the outset:

As a minimum, good advice should be „workable“ or „truly prescriptive“ in taking into account the limited cognitive and other faculties of real people. Individuals must be able to understand and to carry out the advice with the „cognitive technology“ at their command. But being „truly prescriptive“ or „workable“ is not sufficient if we are searching for good advice. Neither is it sufficient that a given workable advice as a matter of fact yielded good results. In the concept of good advice both, procedural elements as well as aspects of material success seem to play a role. For instance, last week’s advice that people born in November should spend their whole savings on a German Lotto ticket

„since the stars were favorable“, was simple and may have yielded a tremendous return on investment for some „peculiar“ individual who followed that advice. But even for the winner the advice was certainly not „good advice” in any meaningful sense of that term. First, there was no systematic relation of the advice to some reasonable theory. Second, ex ante, success could not be predicted with a probability meeting the requirements of reasonable aspiration levels. Third, the advice could not have been generally observed and still have led to a reasonable expectation of success for all addressees of the advice.

What is „reasonable” has not been specified yet. We believe that discussing the concept of „good advice“ will shed some light on the notion of rationality within the concept of „bounded rationality“ and thereby also on what is reasonable. We also believe that the discussion of the self-referential aspects of advice giving in theories of (boundedly) rational behavior will shed some interesting new light on the topics of „self-fulfilling” and „self-refuting prophecy” as traditionally discussed in social science. Like the bank that will indeed crash if the theory that it will crash is commonly believed, a theory’s advice may support itself as good advice if commonly adhered to – or vice versa.

3. The decision environment

From here on, we will refer to the theory of rational behavior under consideration as t. The theory t is a set of descriptive and possibly also of prescriptive „sentences“ or „statements“. Let t apply to a class Γ of games G with strategy profiles S. In the most simple case the theory t assigns a subset of S as (universal) advice to all elements G from the considered class of games Γ. For the sake of specificity and illustration, we shall rely on the following simplifying assumptions about the class of problems considered here:

Let $I:=\{1, 2, ..., n\}$, $n\geq 2$, be the set of players and let Γ denote the class of all games G with

(A.0) $G=(S_1, S_2, ..., S_n; u_1, u_2, ..., u_n)$;

where for all $i\in I$ the

$S_i \neq \emptyset$ are finite strategy sets,

$S := \prod_{i=1}^{n} S_i$ is the set of strategy profiles $s=(s_1, s_2, ..., s_n)$

u_i are mappings $u_i: S \rightarrow \boldsymbol{R}$, $u_i(s) \in \boldsymbol{R}$, which represent the individuals' preferences by a conventional cardinal utility measure.

Assumption:

(A.1) The decision environment can be adequately described as in (A.0).

4. Formal concepts of advice

Due to (A.1) we can restrict ourselves to considering advice φ as a mapping that assigns to each $G \in \Gamma$ a profile of subsets of the strategy sets of G. More specifically, we assume that for all games G from the relevant class Γ the theory t of boundedly rational play specifies for each player i some set of strategies from his strategy set as advice φ_i:

(A.2) $\forall G \in \Gamma, \forall i \in I: \emptyset \neq \varphi(G) \subseteq S_i$.

As opposed to *particular advice* which addresses some non-empty proper subset of the individuals I who are going to play G, let *universal advice* address all players $i \in I$ simultaneously. More specifically,

(D.1) For any $G \in \Gamma$ *particular advice* is a set $\varphi_J(G)$ of strategy profiles $s_J = (s_i)_{i \in J}$ for some subgroup J of the player set I deemed "rationally" eligible by the theory t of (boundedly) rational play of games $G \in \Gamma$ from which the advice is derived.

(D.2) For any $G \in \Gamma$ *universal advice* is the subset $\varphi(G) \subseteq S$ of *strategy profiles* deemed „rationally“ eligible by the theory t of (boundedly) rational play of games $G \in \Gamma$ from which the advice is derived.

We require that universal advice $\varphi(G)$ can be decomposed such that

(A.3) $\forall G \in \Gamma, \forall i \in I: [\, \emptyset \neq \varphi_i(G) \subseteq S_i \wedge \varphi(G) = \prod_{i=1}^{n} \varphi_i(G) \subseteq S]$

Requirement (A.3) is quite strong. For example, in a standard „battle of the sexes” game the advice to select one of the two (strict) equilibria $s^1 = (s_1^1, s_2^1)$ and $s^2 = (s_2^1, s_2^2)$ would be ruled out.

2 / 1	s_2^1	s_2^2
s_1^1	1, 2	0, 0
s_1^2	0, 0	2, 1

Figure 1

Such coordination of specific actions among two players, where the choice of one player must be contingent on the choice made by the other, cannot be accomplished if (A.3) obtains. It would be possible only to give the advice to play one of the strategies leading to an equilibrium. However, this advice would amount to $\varphi(G)=S$.

Let us assume that particular advice for some subgroup of the whole set of individuals can be decomposed likewise. Unless indicated otherwise, the assumptions (A.1-A.3) will be made throughout in this paper.

(D.3) Universal advice is unambiguous on a game G iff $|\varphi(G)|=1$ and it is ambiguous on G iff $|\varphi(G)|>1$.

(D.4) Particular advice is unambiguous for a subset J of all players iff $|\varphi_i(G)|=1$ for every player $i \in J$ and it is ambiguous iff $|\varphi_i(G)|>1$ for some player $i \in J$.

5. Knowing and following advice

Whether for an individual $i \in I$ having received advice $\varphi_i(G)$ it is good policy (not) to act upon the advice depends on circumstances c.

(D.5) If an individual i in a game $G \in \Gamma$ under circumstances c is disposed to act according to $\varphi_i(G)$, we say that she complies with $\varphi_i(G)$.

Compliance as used here does not amount to actual behavior but only to the assent and intentions of the individual. Individuals who are „theory compliant" *intend* or *plan* to act according to the theory. But individuals who are compliant may still deviate from their plans. In particular, individuals may make mistakes in carrying out their intentions.

The most conventional compliance and knowledge assumptions are:

(C) For all $G \in \Gamma$, all players $i \in I$ comply for all circumstances with the advice $\varphi_i(G)$.

(K) For all $G \in \Gamma$, among all players $i \in I$ the (universal) advice $\varphi(G)$ and compliance (C) are common knowledge.

If the advice of theory t and compliance of its addressees are common knowledge, then all individuals know that they receive the same advice $\varphi(G)$ and intend to follow it. Of the known advice $\varphi(G)$ only „part" $\varphi_i(G)$ applies directly to the actions of $i \in I$. Each theory compliant individual i intends to go along with her own advice $\varphi_i(G)$. Of course, since individuals have disjoint strategy sets, their particular advice will be disjoint as well. But if individuals are fully rational and thus do not face any cognitive constraints in reaching inferences, then what individuals know or can know converges among

individuals. Advice is no exception to this. In terms of a strict application of theory t, each and every individual $j \in I$ will reach the same conclusion for each individual $i \in I$. Therefore if, according to the rules of game G, the theory t of how to play it is common knowledge, then the (universal) advice $\varphi(G)$ is commonly known, too.

But this does not imply that players can fully predict other players' behavior. Uncertainty will arise in so-called games of „incomplete information" (see Harsanyi 1967/8). Here players would know the type distribution, but would not know the actual types involved. In such games a closed theory t of rational behavior would fully specify the advice for all types. But since at least some types remain private information, there is uncertainty about which advice would apply. To put it slightly otherwise, which advice of the theory t actually applies is contingent on type, but the conditional advice itself is commonly known if the theory is commonly known.

In classical rational choice modeling „all is known" commonly, since all individuals are in command of the theory of rationality t itself, fully theory compliant and can instantaneously derive all of t's relevant implications. The remaining uncertainties are purely stochastic and can be incorporated in standard ways by expected utility theory.

We can now define – in honor of O. Morgenstern – what may be called the „theory absorption criterion" for an ideal setting:

(D.6) A theory t is absorbable in a class of games, Γ, if for all $G \in \Gamma$ common knowledge of t along with condition (K) does not provide a reason against compliance (C) with t's (universal) advice $\varphi(G)$.

Common knowledge of t relates to its predictive as well as its prescriptive content, whereas (K) concerns t's prescriptive or advice-related aspects. Equilibrium theory (see Cournot 1838; Nash 1951) claims to be „absorbable" in the sense of (D.6): For a class Γ, all players can know a classical theory t of fully rational behavior in any $G \in \Gamma$, can be compliant with its advice, and on the premise of universal compliance will never have good reason to prefer not to follow theory t's advice. In short, if a theory t is absorbable then, if t is

becoming known, this should not be detrimental to the quality of t's predictions and prescriptions as evaluated by the actors themselves.

Ideally a theory t of fully rational behavior should even be self-supporting in the sense that common knowledge of t along with (K) provides a good reason to comply with the theory's advice. Indeed, the existence of unambiguous advice of a commonly known theory t along with (K) provides a reason for intending to follow t's advice. But if advice is ambiguous in the sense of (D.3), it is not clear what it amounts to in behavioral terms (the issue of actually *behaving* rather than merely intending to behave according to t will be addressed in the next section 6.). Then t will in general not be self-supporting since even common knowledge of t does not give us a good reason to believe that its predictions hold good and its prescriptions apply. Likewise, the advice of playing one of the strict equilibria in the battle of the sexes game mentioned before violates (A.3) and therefore does not imply a definite recommendation for individual choices. Theories t leading to vague advice can be absorbable due to the their very vagueness. But in such cases absorbability obviously does not amount to much.

The substantive constraint that rational choice is forward-looking and evaluates all choices in terms of their future causal consequences is relevant to all forms of choice making. It is not specific to interactive situations. The formal requirement that preferences and beliefs fulfill axioms such that they can be represented by appropriate utility and probability measures is not specific to interactive situations either. For such situations the concept of absorbability of the theory is an additional requirement. It constitutes the most central standard specific to interactive situations. It must be met by any a priori theory t explicating what rationality should mean for some class of games Γ.

This standard can be met by theories of full rationality for some classes of games. However, outside classical or extreme rational choice modeling the preceding conclusions about common knowledge as well as the absorbability of theories seem either precarious or inapplicable. It is useful to look at some of the reasons why this is so, to classify and categorize them and then turn again to the issue of theory absorption in a bounded rationality context.

6. Responding to advice

Let us assume for the following that no incomplete information games are included in Γ. Therefore besides common knowledge of the advice, player types are common knowledge. Given these premises, let us scrutinize the relationship between advice and behavior more closely.

6.1. Unique advice and fully rational responses

To begin with, let us make some heroic assumptions: First, the theory t of rational play for the class Γ is such that for all $G \in \Gamma$ the theory t implies definite advice in the following sense:

(A.4) $\forall G \in \Gamma: |\varphi(G)|=1$.

Second, all players do *not only intend* to act according to what they accept in theory (C),but beyond (K) it is also common knowledge among them that they *do in fact* act that way (with probability 1). Knowing the theory t and the advice $\varphi(G)$, players also know what *behavior* to expect from other players in G.

If advice $(\varphi_1(G), \varphi_2(G), ...,\varphi_n(G))=(s_1, s_2, ..., s_n) \in S$ is a singleton – that is, if (A.4) applies – it is fully specified what players who fulfill (C) intend to do. All act upon the universal advice received, know the content of that advice and know exactly what it means in behavioral terms to follow the advice $\varphi(G)$ of theory t.

Assuming perfect rationality and decomposing $\varphi(G)$ as $\varphi(G)= (\varphi_i(G), \varphi_{-i}(G))$ and $(\varphi_1(G), \varphi_2(G), ...,\varphi_{i-1}(G), s_i ,\varphi_{i+1}(G), ..., \varphi_{n-1}(G), \varphi_n(G))$, respectively, the best reply requirement amounts to the demand

(BR) $\forall G \in \Gamma, \forall i \in I, \forall s_i \in S_i: u_i(\varphi_i(G), \varphi_{-i}(G)) \geq u_i(s_i, \varphi_{-i}(G))$.

Along with (C) and (K), condition (BR) amounts to the classical definition of an equilibrium. Better advice cannot be given to any individual if all individuals do in fact behave according to the advice.

Given the assumption that behavior coincides with advice, what may be called „behavioral equilibrium" exists if „advice equilibrium" does and vice versa. That this is so is assumed by practically all theories of fully rational behavior. But exactly at this point we need to bridge the gap between theory and actual behavior. There must be a *causal* link between the theory t's *advice* and the *behavior* of individuals. Without a causal link between a theory t's advice – as perceived by its addressees – it would be pure magic should behavior correspond to advice. But can such a causal link legitimately be assumed to exist and if so, how does it work? To answer this question by simply assuming that advice and behavior coincide is not good enough in theories of boundedly rational behavior. For, from the empirical point of view of merely boundedly rational behavior, it is by no means obvious that advice and behavior coincide.

Of course, individuals could violate (C) and behave intentionally contrary to advice. But this is not the essential point. The essential point is that even if individuals intend to be theory compliant, they might still make mistakes and, as we shall argue below, they might simply be unable to follow a theory t's best advice. This problem is particularly relevant if the theory t, in deriving its advice, is presupposing fully rational individual behavior according to (BR). – It is exactly here that some of the more interesting topics concerning the relationship between fully and boundedly rational *behavior* are located.

6.2 Unique advice and boundedly rational behavior

If we think of real individuals with their several „real" limitations, then even if (A.4) is fulfilled and advice unambiguous, the reasoning leading to the selection of an equilibrium under theory compliant best reply behavior will apply merely in very simple cases (for a more extended discussion of „complexity", see 6.3 below). For instance, if in a two by two game each decision maker is advised to use a strictly dominant strategy, and if the performance of the advice singles out a payoff dominant equilibrium, then individuals should be expected to follow the advice in practice. Moreover, their behavior will amount to best response

behavior in such simple cases, regardless of the assumption that they face emotional and cognitive constraints in their choice making.

However, this coincidence of the behavior of boundedly rational individuals with that of fully rational individuals does not go a long way. Consider the following simple game

2 / 1	s_2^1	s_2^2
s_1^1	101, 101	0, 102
s_1^2	102, 0	1, 1

Figure 2

Assume that the game of figure 2 is introduced to both players while both are present and aware of the presence of each other. Through this publicness of the introductory event, the players acquire a kind of „working common knowledge“ of the game. They know that the other must know it etc. Moreover, assume that there is no „cheap talk“ among the players during the public event. Finally, after the matrix has been introduced, the players are separated such that no communication is possible between them when they are actually making their choices.

If we focus exclusively on pure strategies in the game of figure 2, we have three payoff-undominated strategy combinations that are not in equilibrium and a single strict, yet payoff-dominated, equilibrium $s^2=(s_2^1 , s_2^2)$ in dominant strategies. If for such a game, G, advice $\varphi_i(G)$ would suggest to each player the choice of her dominant strategy, then fully rational players would intend to act accordingly. In a „full rationality setting“, this would clearly be the only acceptable outcome. But it is not an intuitively appealing outcome for human individuals and their common sense.

Assume therefore that players are not fully rational but willing to go against the dominance principle under certain circumstances. If the commonly known theory of boundedly rational play t that allows for playing strictly dominated strategies would entail the advice $\varphi(G)= (s_1^1 , s_2^1)$ for the game, G, of figure 2, then this advice could be followed without self-refutation among *satisficing* players (see Simon 1955). If such players endorse aspiration levels of, say, 100 each, none would have reason not to comply with the advice. The aspiration level will in fact be met only if the co-player is following the advice $\varphi(G)=(s_1^1 , s_2^1)$. In that sense the theory of boundedly rational behavior induces genuine strategic interdependence where a theory of full rationality would have individuals look only at the payoffs of their own dominant strategy.

The individual has good reason to follow the advice $\varphi(G)=(s_1^1 , s_2^1)$ only if her theory t *descriptively predicts* that other individuals will endorse appropriate aspiration levels and therefore follow the advice. This provides a first glimpse of how descriptive and normative components are interwoven within a bounded rather than a full rationality approach. A player with a fixed aspiration level of, say, 100, will be satisfied with the result of observing the theory t iff, *as a matter of fact, her co-player* goes along with the theory, too. Her co-player must start from a compatible aspiration level as well etc.

The interaction between predictive and evaluative aspects of advice generating theories goes much further, though. In the full rationality context, the normative theory t could simply be used to generate the predictions on which its own normative prescriptions rest. Fully rational individuals would simply observe the prescriptions of t and then transform them into predictions. However, now something needs to be said about how individuals are actually dealing with theories. One obvious aspect of this emerges if we take into account the limited ability of players to cope with the complexity of strategic advice. In particular, a theory t of boundedly rational behavior may predict that some advice φ may be too complex to be followed. Then the assumption that the addressees of advice will behave according to the advice no longer holds. Let us briefly look at the

issue of cognitive constraints first and then turn to emotional constraints before facing the intricate problem of ambiguity.

6.3 Cognitive limitations and complexity

Conventionally, the issue of complexity is illustrated by the example of chess playing. In chess, conceivably players might learn a „choreography“ of the moves for a specific way of playing the game based on the theory of „rational chess play“. The play, in the sense of two sequences of moves – one for each player – is the result of strategies. However, the play is not the same thing as a strategy profile. Even if the strategic advice $\varphi(G)=(\varphi_1(G), \varphi_2(G))$ of the theory were – as assumed here for the sake of argument – a singleton, there would be no way for human players to memorize full chess strategies $\varphi_i(G) \in S_i$, i=1, 2. Real players – and real machines as well – must rely on simplifying advice rules at some point. Bad chess players will rely on very simple advice rules, good players presumably on more complicated ones. In any event, there will be no complete, complex casuistic, but rather a set of relatively simple advice rules employed by the players with some complementary casuistics.

Players will apply advice rules, but in each situation they will, *as they go along, construe* the decision situation to which they apply the advice rules. Moreover, unlike in a computer chess program, there will always be implicit or tacit instances of violating the advice rules that cannot be specified in advance (on implicit knowledge, see Polanyi 1962). In a specific situation which is unforeseen by the boundedly rational player himself (and perhaps unforeseeable altogether), there may be „a reason to deviate” unspecified by the advice rules, but fully plausible to the human mind (links with this range from the antique „topoi“ – Aristotle – to the modern „unless“ clauses in expert systems).

Responding to the need to rely on advice rules, complexity may be the origin of predictable behavior (as e.g. claimed in Heiner 1983), but it clearly remains a source of unpredictability as well. Complexity in the relevant sense may emerge easily even in extremely simple games if we allow for ascending levels reflection. The so-called guessing game may serve as an illustration here. In a

very simple version of this game, n≥2 individuals j=1, 2, ..., n must name a number from the set {0, 1, 2, ..., 100}. The guesses s_j of all players are summed up and averaged to yield half of the average according to $(\sum_{j=1}^{n} s_j)/2n$. The player i wins if his guess turns out to be correct, that is, if $s_i=(\sum_{j=1}^{n} s_j)/2n$.

Since the maximum of all guesses is determined by n*100, half of the average bids cannot exceed 50. Therefore, all rational players i will exclude all bids s_i >50. In view of this, any rational individual should come to the conclusion that no rational individual will bid more than 50. However, then the maximal sum of bids is clearly n*50, and nobody should guess more than 25. In turn, the maximal sum becomes n*25 etc.

The preceding line of argument clearly depends on the presumption that (C), (K) and (BR) apply, and hence every player assumes the others follow the „logically" derived advice. But an individual who starts to reflect on whether or not advice once given will be followed, may eventually become entangled in such a complex web of thoughts that he could no longer handle the complexity. This inability – as has been observed frequently – may look like adopting a mixed strategy: A boundedly rational player who starts engaging the complexities of strategic interaction may become unpredictable to himself and others simply because she cannot draw all conclusions from her own model of the situation. This is an interesting justification for a behavioral and not merely expectational interpretation of mixed strategies. But it should be noted, too, that this view of mixed strategies applies only in a bounded rationality and not in a perfect rationality framework.

6.4 Emotional limitations and weakness of the will

Cognitive complexity is not the only relevant influence. Thinking of the chain store paradox or centipede games (see Selten 1978 and also Rosenthal 1981) one can well imagine that real human players might understand and accept the logic of backward induction. At the same time, due to motivational limitations they

may not want to follow the unique advice of the theory that they fully understand and know to apply in simple situations like centipede games. For Selten (1990) such deviations from the precepts of backward induction form a clear instance of boundedly rational as opposed to fully rational behavior. But others have argued that a concept of full rationality that implies backward induction is inadequate (e.g. McClennen 1990). According to this view, some modifications of the present concept of full rationality must be made if a reflective equilibrium in explicating the concept of rationality is to be reached (see section 1 above).

It may well be that some far-reaching modifications of the standard concept of full rationality are necessary. However, it seems quite doubtful that the arguments presented so far are sufficient to show this. At least some cautionary remarks seem appropriate. First, the argument cannot plausibly be restricted to sequential games and backward induction. As the examples of the guessing game and the game of figure 2 show, one-off games pose problems that would require similar modifications. Second, the alleged need for a modification of the rationality concept may in the last resort be based entirely on emotional reactions. Third, the modified concept of rationality that eventually emerges if emotional reactions are factored in may be one of bounded rather than full rationality.

All the problems mentioned so far will arise even if advice is universal and unambiguous. Ambiguous advice raises additional problems. Some we will now briefly address and then come back to the more general methodological points of explicating the concept of rationality adequately.

6.5 Ambiguity of advice

Clearly, if for a class of games Γ the theory of rational play implies unambiguous advice and if the singleton set $\varphi(G)$ is assumed to be common knowledge, then in every game $G \in \Gamma$ an equilibrium should be selected. This holds good if players expect each other to comply with it, to act in fact upon it and then react optimally to the advice (see Aumann and Brandenburger 1995).

Even critics of the exclusive focus on equilibria in traditional noncooperative game theory would not deny this (see, e.g., Sugden 1991). But they would still insist that the assumption that for a class of games Γ the theory of rational play implies unambiguous advice is farfetched unless the class Γ is a very restricted one (in particular, if unambiguous advice has to meet certain plausible standards of consistency that would rule out that the theory t could by its advice select among equilibria). They maintain that there will always be games for which any reasonable theory t implies an advice vector $|\varphi(G)|>1$.

If $|\varphi(G)|>1$, then there is at least one player $i\in I$ for whom $|\varphi_i(G)|>1$. Even if the other players knew the advice $\varphi(G)$ and knew that all players intended to act according to the advice, none of the players would know which action the player i should and therefore would take. There would be at least two acts compliant with the advice $\varphi(G)$ of the theory t.

One might represent this uncertainty by expectational probability distributions p_{ji} endorsed by the individuals $j\neq i$, if $|\varphi_i(G)|>1$ obtains. Of course, $p_{ki}\neq p_{ji}$ is possible. In traditional game theoretic reasoning all $k, j \in I$ would have the same views $p_{ki}=p_{ji}$ in common. As deductive theory has it, such common views would emerge as the result of some procedure or other applied to common priors obtaining on some ultimate level of analysis. However, it is hardly conceivable that the sophisticated methods of advanced formal game theory (as, e.g., in Harsanyi and Selten 1988) would represent what is actually going on in the minds of boundedly rational players. Boundedly rational players can manage only very restricted problems. They proceed inductively rather than deductively, looking selectively at certain aspects of the game, construing the action situation to fit their limited cognitive capabilities and eventually applying some rules of thumb rather than optimal decision rules to derive an advice as to how to act.

But instead of dealing with such problems in the abstract, let us again discuss a specific example. Consider

table of the game					criteria for player 1		
s_1 \ s_2	s_2^1	s_2^2	s_2^3		$\min u_1$	$\max u_1$	$(1/3)\Sigma u_1$
s_1^1	4,4	3,2	3,1		3	4	(1/3)10
s_1^2	2,3	5,5	4,1		2	5	(1/3)11
s_1^3	1,3	1,4	6,6		1	6	(1/3)8

Figure 3

All equilibria in pure strategies are strict: (s_1^1, s_2^1), (s_1^2, s_2^2), (s_1^3, s_2^3). But can we restrict advice to such pure strategies?

Since in the end each player must choose to do something, and all acts that can be „done“ are already listed in the table, mixed strategies cannot be „chosen“ in the proper sense of the term. The advice to randomize strategies is viable only if there is such an option. In that case, the choice of the relevant „random device“ amounts to the choice of an additional strategic option.

If present, the option to randomize should explicitly show up as an additional strategic choice besides the other choices given in the table. If, for example, in the preceding game player 1 could indeed choose to play a mixed strategy with probability parameters (p_1^1, p_1^2, p_1^3) ,then this strategy should show up as an additional option $s_1^4 := p_1^1 s_1^1 + p_1^2 s_1^2 + p_1^3 s_1^3$. Physically ,this option might emerge from having access to dice or a coin to be thrown in an appropriately defined random experiment, or even the observation of some „random event” in the environment (where we neglect the difficulty that without commitment

power the decision to choose according to the random result would still have to be made).

If options to randomize do not exist, then players cannot *choose* „mixed“-strategies. Players may nevertheless *intend* to follow the advice to mix if their theory t of rational play suggests that. However, for boundedly rational individuals, intentions and actual behavior are two different things. In fact, experimental evidence shows that players with limited memory capacity who intend to randomize their strategies according to some given distribution are hard-pressed and prone to show severely biased behavior (see Kareev 1992).

As opposed to „behavioral” there is „perceived” strategy mixing. A player k who is uncertain about the behavior of her co-player j may express her uncertainty about his choices formally by treating him as if he were playing a mixed strategy. Such mixed strategies are like „beauty in the eye of the beholder”. The advice φ to „choose” such a „mixed strategy”, that is, to *behave* accordingly, does not make much sense. As far as the choice of strategies is concerned, only a behavioral and not a perceptual interpretation seems appropriate. However, only pure strategies – including, though, the choice of a random mechanism or random device – can be *chosen* in the proper sense of that term.

If only pure strategies can be chosen, then in the example of figure 3 advice must single out one or several pure strategies. In view of (A.3) ,the advice „select an equilibrium” does not do. But the advice „select one of the strategies that lead to an equilibrium!“ amounts to

$$\varphi(G)=(\varphi_1(G), \varphi_2(G)) =(\{s_1^1 , s_1^2 , s_1^3 \}, \{ s_2^1 , s_2^2 , s_2^3 \}).$$

Since this advice does not exclude any option, it is in fact without content. Even if φ(G) was made known in a public event, this fact would not give rise to any more definite expectations of the players.

The ambiguity of the advice of choosing only strategies that are part of some equilibrium may not be of much practical relevance. Boundedly rational

decision makers will not think in terms of equilibria anyway. They will use some rule of thumb or other. But such rules of thumb as conventionally used by boundedly rational decision makers may, and in all likelihood will, lead to contradictory advice in the game of figure 3 as well. Therefore, relying on boundedly rational behavior may not overcome the difficulty.

To give just one example of how each of the strict equilibria may be singled out by a plausible rule of thumb (for a recent experimental study illustrating such, in their terminology, „pessimistic", „naive" and „optimistic" rules of thumb, see Costa-Gomez et al. unpub., 2000):

Strategies s_i^1 , i=1, 2, yield the largest worst-case payoff. These strategies are singled out by a maxi-min rule of thumb.

Strategies s_i^2 , i=1, 2, yield the highest expectation if each player forms expectations about the co-player's behavior in the somewhat Laplacean way of ascribing an equal likelihood to each choice of the co-player. These strategies are singled out by the advice of maximizing expected payoff under an equal likelihood assumption for situations where there is no special reason to favor any.

Strategies s_i^3 , i=1, 2, yield the largest possible outcome. These strategies are singled out by a maxi-max rule of thumb.

If decision makers know that the three preceding rules of thumb are plausible candidates and at the same time do not know which one is in fact used, they cannot draw any further conclusions from this information. As long as none of the rules of thumb is privileged over the others, it does not help individuals to mutually emulate their boundedly rational ways of decision making. Putting themselves in each other's shoes, they still end up with the advice vector $\varphi(G)=(\varphi_1(G), \varphi_2(G)) = (\{s_1^1, s_1^2, s_1^3\}, \{s_2^1, s_2^2, s_2^3\})$ which does not exclude anything from consideration.

In a sense, such null advice is just an extreme case of ambiguous advice. Players need sources other than the theory t and its derived advice φ(G) for predicting choices. Nonunique advice will leave players somewhat at a loss *even if they assume that empirically individuals know the theory t and intend to follow its advice*. Expected behavior must be constrained by some facts or on empirical grounds (e. g. „conventions") if individuals are to receive more definite advice from the theory t (on the evolution of conventions, see the experimental studies by Berninghaus and Ehrhart, 1998 and more generally Young 1993). This reduction of ambiguity is „inductive" rather than deductive and clearly not deduced from the theory t that specifies the rationality concept for the context under consideration.

7. Causality and absorbability of advice

Both, theories of fully rational behavior as well as theories of purely adaptive behavior, avoid the problem of explicitly modeling the causal influence of a theory t and its advice on actual behavior. Theories of fully rational behavior not only assume that (C) and (K) are fulfilled, but also that behavior is fully compliant with t. Therefore, theories of fully rational behavior can predict behavior as being coincident with their own advice. The advice itself is derived on the *assumption* that prescribed and predicted behavior amount to the same among fully rational beings. This full rationality approach is coherent, but one should not forget that it is based on assumptions rather than on empirical laws or findings. Likewise, a purely adaptive „non-rational" approach is based on assumptions that abstract away the higher faculties of the mind, in particular those that enable the understanding of theories. Due to this abstraction, theories of non-rational, purely adaptive behavior eliminate the problem of modeling the causal influence of the theory t *of* behavior *on* behavior. Starting from the premise that there is no causal influence of the theory t on behavior at all, t can treat behavior parametrically and predict it without referring to t and its advice.

Neglecting the causal influence of a theory t of rational behavior on behavior may be adequate in some extreme cases. But in many non-extreme situations it

may distort the facts. Moreover, there is no good reason to assume that intermediate cases can be treated by a kind of convex-combination of the treatment of the extreme cases. What is needed is some theory that fills the middle ground. This theory should incorporate some of the idealizations characteristic of theories of fully rational behavior and at the same time observe some essential facts and constraints influencing actual choice making. As far as such theories are concerned Tietz (1992) has coined the concept of a „semi-normative theory". A semi-normative theory t takes into account some stylized facts of actual human behavior considered rational according to some convention or other. At the same time, a semi-normative theory tries to adhere to some idealization of behavior. In particular when deriving theoretical advice, it is assumed that individuals are aware of the presence of other individuals and their reasoning in ways akin to the assumptions (C) and (K), as conventionally used in theories of full rationality.

A semi-normative theory t assumes that individuals have only a limited capacity for behaving according to t's best advice. But they can be influenced by the theory, and therefore the theory t's advice must take into account the theory's own causal effects on behavior. For example, if in the game of figure 3 player 2 comes to the conclusion that theory t provides player 1 with the advice to play s_1^2, then player 2 would not want to use all her strategies with equal probability but rather single out s_2^2. A theory t that would suggest that player 2 play each of her strategies with equal probability, while player 1 receives the advice to play his second strategy with probability 1, would seem incoherent not only among fully but also boundedly rational individuals.

For semi-normative theories t as for theories of ideal rationality, the problem of absorption of the theory t, (D.6), does play a role. As opposed to the counter-factual assumptions of normative theories of full rationality in the world of semi-normative rationality, the causal effect of the theory itself must be viewed as subject to factual laws. Therefore, in formulating a convincing semi-normative theory t, we need to address the problem of how the predictive content of t relates to t's prescriptive content. If in a game G some advice $\varphi(G)$

is given, then all depends on what players make of it. This is an empirical issue to be addressed by the theory t of behavior used for deriving the advice. For instance player 2 must ask himself what he assumes player 1 to make of the advice $\varphi(G)$; where $\varphi(G)$ is possibly derived from the same theory t that is used to predict the behavior of using the advice. Even among boundedly rational players the thinking of player 1 may include his thoughts about the thought processes of player 2. Even a merely boundedly rational player 1 might ask questions like the following: What will player 2 make of the advice $\varphi_2(G)$? If $|\varphi_2(G)|>1$, then player 1 may ask how likely it is that player 2 upon receiving the advice $\varphi_2(G)$ will choose any of the alternatives in $\varphi_2(G)$.

What is needed – including the case of unique advice – is an *empirical* theory of the process in which choices are in fact made and how advice once developed is itself processed in view of such an empirical theory of decision processes. This search is complicated by certain self-referential elements since the theory of (boundedly) rational behavior from which advice is derived can – and in general will – itself causally influence the process. The theory t is *one* of the informational inputs into the process of boundedly rational decision making. But, whatever we may assume about the role of the theory and its advice, in the last resort *the effects of the theory and its advice are causal rather than logical.* The theory must be assessed in causal rather than logical terms.

Tietz (see again 1992) suggests that semi-normative theories t be tested by inquiring to what extent such a theory t could become known among all boundedly rational individuals and be followed by them without self-refutation. As in the standard case of theories of fully rational behavior (D.6), all individuals should be in a position to assume that all other individuals follow t and still have no incentive to deviate from the theory's advice. This „test" looks quite similar to a standard equilibrium condition. This familiarity may seem advantageous at first glance, but not on closer examination. For it is based on the *assumption* that *all* individuals follow t, rather than on an empirical theory of how many individuals can in fact be expected to follow t and to what extent. The challenge is to formulate a semi-normative theory t of boundedly rational behavior that contains an empirical prediction about the extent to which it will

itself be accepted by players who are aware of it and causally influence their behavior. Addressing this issue may bring us closer, too, to a solution of the conceptual problem of what makes the behavior predicted by an explanatory valid theory t of boundedly rational behavior a „rational“ rather than a merely „predicted“ behavior.

8. Conclusions

In view of the desired continuity with common intuitions about what is rational, we may want to modify the rationality concept. Certain kinds of behavior that are not captured by the classical concept of rationality can then be classified as rational. To put it slightly otherwise, behavior that is deemed non-rational by concepts of full rationality, but seems rational to common sense can be treated as fully rational then. At the same time, the theory will not imply advice such as never to cooperate in a centipede game. It will therefore itself not be potentially subversive for proclivities to cooperate if co-operation is – at least potentially – highly profitable without imposing high risks.

Still, the insistence that one of the emerging non-extreme explications of the concept of rationality should be taken as the ideal type or the limiting case of full rationality raises doubts for the very reason that it is not an extreme case. In particular, if we believe that human behavior is characterized by the faculty (an ability rather than a vice) of acting opportunistically, then this tends to drive us to extremes. Along with the assumption that individuals are able to engage in increasing levels of reflection it almost implies the principle of backward induction. Beings who command the faculty of engaging in forward-looking opportunistically rational choice, aiming at states that are to their individual benefit, will be advised to follow the logic of choosing the best action in terms of the anticipated future (causal) consequences. If we take this seriously, then the „rational” in „boundedly rational behavior” will always be „imperfect”.

On the other hand, if we look at behavior as being purely adaptive, then theories and the advice derived from them will play no role. Purely adaptive views of human behavior are questionable precisely because they neglect their own

impact on behavior. Best reply dynamics, however sophisticated they may be, cannot pass the test demanding of them to take into account the impact of theories of such dynamics on the dynamics themselves. As the example of the so-called „self-fulfilling prophecy" shows, social scientists have been aware of the effect of „theoretical" predictions on behavior for a long time. But like their game theoretic peers, they did not contribute much to reaching the middle ground between theories of perfect rationality and theories of no-rationality. We need to specify the meaning of „rational" in theories of boundedly rational behavior and at the same time take into account the facts of behavioral adaptation.

References

Aumann, R., and Brandenburger, A. 1995. Epistemic Conditions for Nash Equilibrium. *Econometrica* Vol. 63(5), 1161-1180.

Berninghaus, S. and Ehrhart, K.M., 1998. Time Horizon and Equilibrium Selection in Tacit Coordination Games: Experimental Results, *Journal of Economic Behavior and Organization*, Vol. 37, no. 2, pp 231-249

Carnap, R. 1956. *Meaning and Necessity*. Chicago: University of Chicago Press.

Costa-Gomes, M., Crawford, V. P. and Broseta, B. unpub., 2000. Cognition and Behavior in Normal-Form Games: An Experimental Study, University of California, San Diego, Discussion Paper 2000-0R, July 2000.

Cournot, A. 1838. *Recherches sur les principes mathématiques de la théorie des richesses*. Paris: L. Hachette.

Harsanyi, J. C. and R. Selten 1988. *A General Theory of Equilibrium Selection in Games*. Cambridge, Mass.: MIT Press.

Heiner, R. 1983. The Origin of Predictable Behavior. *American Economic Review* 73/4: 560 ff.

Kareev, Y. 1992. Not That Bad After All: Generation of Random Sequences, *Journal of Experimental Psychology, Vol. 18, No. 4,* 1189-1194.

McClennen, E. F. 1990. *Rationality and Dynamic Choice - Foundational Explorations*. New York et al.: Cambridge University Press.

Nash, J. F. 1951. Noncooperative Games. *Annals of Mathematics* 54: 289-295.

Polanyi, M. 1962. *Personal Knowledge.* Chicago: University of Chicago Press.

Rosenthal, R. 1981. Games of Perfect Information, Predatory Pricing, and the Chain Store Paradox. *Journal of Economic Theory* 25: 92-100.

Selten, R. 1978. The Chain Store Paradox. *Theory and Decision* 9: 127-159.

Selten, R. 1990. *Some Remarks on Bounded Rationality*. Bonn: Sonderforschungsbereich 303 „Information und die Koordination wirtschaftlicher Aktivitäten“.

Simon, H. A. 1957. *Models of Man*. New York: John Wiley and Sons.

Sugden, R. 1991. Rational Choice: A Survey of Contributions from Economics and Philosophy. *The Economic Journal* 101(July): 751-785.

Tietz, R. Semi-Normative Theories Based on Bounded Rationality. *Journal of Economic Psychology 13*: 297-314.

Young, H. P. 1993. The Evolution of Conventions. *Econometrica* 69: 57-84.

We would like to express our gratitude to Adelheid Baker for improving our English style; of course, the conventional disclaimer applies. W.G.,H.K.